!!!Kasten Bioinformatik

@a:Matthias Lange, Andreas Stephanik

@kt:Datenbanken und Tools der Bioinformatik im World Wide Web

@$:Die in der molekularen Biologie gewonnenen experimentellen Daten werden innerhalb öffentlicher und kommerzieller Forschungsprojekte systematisch erfasst und in entsprechenden Datenbanken gesammelt. Aktuelle Erhebungen systematisieren allein 300 begutachtete Datenbanken (http://www3.oup.co.uk/nar/db2001), die im Internet für die Forschung frei verfügbar sind. Diese Daten umfassen Informationen zu den drei heutigen Schwerpunkten der Bioinformatik: Sequenzanalyse, Proteindesign und Metabolic Engineering. Analog dazu lassen sich entsprechende Vertreter für Tools und Datenbanken einordnen. Oftmals bilden beide Komponenten ein Gesamtsystem.

@z:Sequenzanalyse

@$:Die maschinelle Sequenzierung und das damit verbundene exponentielle Anwachsen der Datenbestände macht die elektronische Datenerfassung und -verwaltung erforderlich. Daneben wurden zahlreiche Algorithmen entwickelt, mit deren Hilfe die gezielte Analyse dieser Datenbestände ermöglicht wird.

@z:Proteindesign

@$:Beim Proteindesign erzeugt man schon seit einigen Jahren unter anderem mit Hilfe der CAD-Technologie wissenschaftlich fundierte Darstellungen von Molekülen. Darüber hinaus veranschaulichen Computersimulationen auch das dynamische Verhalten biochemischer Moleküle. Wissenschaftler können so die komplexe Struktur und Funktion vieler Proteine analysieren, um diese zu modifizieren oder völlig neue Proteine zu entwickeln.

@z:Metabolic Engineering

@$:Auf der Basis der vorhandenen molekularen Datenbestände werden komplexe Informationssysteme implementiert, die den integrativen Zugriff auf diese Datenbestände ermöglichen und die Modellierung und Simulation von metabolischen Prozessen erlauben. Dadurch wird ein Engineering molekulargenetischer Netzwerke ermöglicht.

@z:Herausforderungen für eine praktikable Dateninfrastruktur

@$:Das WWW und die darin angebotenen Datenbanken der Molekularbiologie und dazugehörige Analyseprogramme haben sich zu einem mächtigen, vielfach genutzten Werkzeug entwickelt. Derzeit wird intensiv diskutiert, welche Methoden zur
dauerhaften Datenhaltung und -präsentation verwendet werden sollten. Am vielversprechendsten scheint der Einsatz von Datenbank-Management-Systemen, unter anderem wegen der Qualität der Anfragemöglichkeiten, des indexierten Datenzugriffs, der Anbindung von externen Werkzeugen und der Anfrageverteilung.

Bisher sind ein großer Teil der frei verfügbaren Datenbanken evolutionär gewachsene Legacy-Datenbanken, die auf einfach strukturierten Dateien (flat files) basieren. Aus diesem Grund ist es notwendig, diesen Datenbestand durch die Verwendung von Methoden der Datenextraktion zu nutzen. Eine verbreitete Methode hierzu ist das Parsen dieser flat files. Dafür sind für jede dieser Datenbanken spezifische Regeln zum Parsen zu entwickeln.

Neben den individuellen Umsetzungen gibt es Bestrebungen, ein einheitliches

Format für die Dateien zu verwenden. So gibt es zum Beispiel das Sequence Retrieval System (http://srs6.ebi.ac.uk), das verschiedene Datenbanken mittels eines allgemeinen Datenextraktionsansatzes und durch ein WWW-Frontend bzw. APIs homogen zugänglich macht. Im Bereich der Schnittstellen zu den Daten
existieren verschiedene Ansätze, die sehr systemabhängig sind. Oftmals bilden nicht standardisierte HTML-Formulare in Korrelation mit serverseitigen Methoden zur Erzeugung dynamischer HTML-Dokumente die einzige öffentliche Schnittstelle zum Datenbestand. Dazu gehört der Einsatz von Shellskripten, CGI/Perl, PHP oder ASP. Die Bereitstellung von deklarativen und komplexen Anfragesprachen für ein System wie zum Beispiel SQL ist eher die Ausnahme.

Neben diesen Heterogenitäten bei der Anfrage ist auch die Anbindung von Anwendungen wie Simulations- oder Analysewerkzeugen an eine Datenbank eher problematisch. Das manifestiert sich oftmals in der mangelhaften Unterstützung von Zugriffs- und Programmierschnittstellen, wie ODBC, JDBC, RMI oder CORBA. Es existieren aber konkrete Bemühungen, CORBA als systemübergreifende Schnittstelle einzusetzen. Für JAVA-Anwendungen wird auch zunehmend das von SUN entwickelte RMI verwendet.

Die öffentliche und kommerzielle Forschung arbeitet zurzeit daran, diese Probleme zu überwinden. Einen Schwerpunkt bildet die Anwendung von Techniken zur Integration heterogener Datenbanken. Das Szenario einer Interaktion mit einer Menge heterogener molekularbiologischer Datenbanken, um den Datenbestand zu sichten, Informationen zu gewinnen, komplexe Anfragen zu formulieren oder Daten zu analysieren, stützt sich zur Zeit auf vier technische Umsetzungen, für die hier jeweils nur Beispiele genannt werden können:

Hypertextnavigation: KEGG (http://www.genome.ad.jp/kegg/kegg2.html),

Data Warehouse: SRS (http://srs6.ebi.ac.uk), BioRS (http://www.biomax.de), HUSAR (http://mbp-ultra3.embnet.dkfz-heidelberg.de/w2h),

Multidatenbank Anfragesprachen: BioKleisli/TAMBIS
(http://img.cs.man.ac.uk/tambis) und

Hybride Ansätze: MARGBench
(http://edradour.cs.uni-magdeburg.de/iti_bm/marg). (anm)

@$:<I>Die Autoren sind wissenschaftliche Mitarbeiter der Arbeitsgruppe Bioinformatik/Medizinische Informatik am Institut für Technische und Betriebliche Informationssysteme der Otto-von-Guericke-Universität Magdeburg.<I>

@ülit:Weitere Linksammlungen

@lit:
http://www-bm.cs.uni-magdeburg.de/iti_bm/marg/dataacquisition/data_sources.html

http://bioinfo.life.nthu.edu.tw/links.htm

http://www.rcsb.org/pdb/links.html

http://restools.sdsc.edu/

!!!Kasten Bioinformatik Ende